Supplementary Material for BEVHeight: A Robust Framework for Vision-based Roadside 3D Object Detection

Lei Yang[1]*, Kaicheng Yu[2], Tao Tang[3], Jun Li[1], Kun Yuan[4], Li Wang[1], Xinyu Zhang[1]†, Peng Chen[2]
[1]State Key Laboratory of Automotive Safety and Energy, Tsinghua University
[2]Autonomous Driving Lab, Alibaba Group; [3]Shenzhen Campus, Sun Yat-sen University
[4]Center for Machine Learning Research, Peking University
{yanglei20@mails, lijun1958@, xyzhang@, wangli_thu@mail}.tsinghua.edu.cn
{kaicheng.yu.yt, trent.tangtao}@gmail.com; kunyuan@pku.edu.cn; yuanshang.cp@alibaba-inc.com

A. Appendix

A.1. Broader Impacts

Our work aims to develop a vision-based 3D object detection approach for roadside perception. The proposed method may produce inaccurate predictions, leading to incorrect decision-making for cooperative autonomous vehicles and potential traffic accidents. Furthermore, we propose a new perspective of leveraging height estimation to solve PV-BEV transformation, facilitating a high-performance and robust vision-centric BEV perception framework. Although considerable progress has been made with our proposed height net and height-based 2D-3D projection module, we believe it is worth further exploring how to combine height and depth estimations to extend to autonomous driving scenarios.

A.2. Contributions

Theoretically, our proposed height-based pipeline entails: i) representation agnostic to distance, as visualized in Fig. 1, ii) friendly prediction owing to centralized distribution as displayed in Fig. 2, iii) robustness against extrinsic disturbance as illustrated in Fig. 3. Technically, we design a novel HeightNet and the projection module with less computational cost. Experimentally, experiments on various datasets and multiple depth-based detectors show the superiority of our method in both accuracy and latency.

A.3. Latency

As shown in Tab. 6, we benchmark the runtime of BEVHeight and BEVDepth. With an image size of 864x1536, BEVDepth runs at 14.7 FPS with a latency of 68ms, while ours runs at 16.1 FPS with 62ms, which is around 5% faster. It is due to the depth range (1∼104m) being much larger than height (-1∼1m), thus ours has 90 height bins that less than 206 depth ones, leading to a slimmer regression head and fewer pseudo points for voxel pooling. It evidences the superiority of predicting height instead of depth and advocates the efficiency of our method.

Table 6. **Latency of BEVHeight and BEVDepth.**

Methods	Backbone	Range	Number of bins	Latency (ms)	FPS
BEVDepth [3]	R50	1 - 104m	206	82	12.2
BEVHeight	R50	-1 - 1m	90	77	13.0
BEVDepth [3]	R101	1 - 104m	206	68	14.7
BEVHeight	R101	-1 - 1m	90	62	16.1

Measured on a V100 GPU. Image shape 864×1536.

A.4. Dynamic Discretization

The height discretization can be performed with uniform discretization (UD) with a fixed bin size, spacing-increasing discretization (SID) [1] with increasing bin sizes in logspace, linear-increasing discretization (LID) [5]and our proposed dynamic-increasing discretization (DID) with adjustable bin sizes. The above four height discretization techniques are visualized in Fig. 8. Following DID strategy, the distribution of height bins can be dynamically adjusted with different hyper-parameter α.

Experiments in Tab. 7 show the detection accuracy improvement 0.3% - 1.5% when our dynamic discretization is applied instead of uniform discretization(UD). The performance when hype-parameter α is set to 2.0 suppresses that of 1.5 in most cases, which signifies that hype-parameter α is necessary to achieve the most appropriate discretization.

A.5. Analysis on Point Cloud Supervision.

To verify the effectiveness of point cloud supervision in roadside scenes, we conduct ablation experiments on both BEVDepth [3] and our method. As shown in Tab. 8, BEVDepth with point cloud supervision is slightly lower

*Work done during an internship at DAMO Academy, Alibaba Group.
†Corresponding Author.

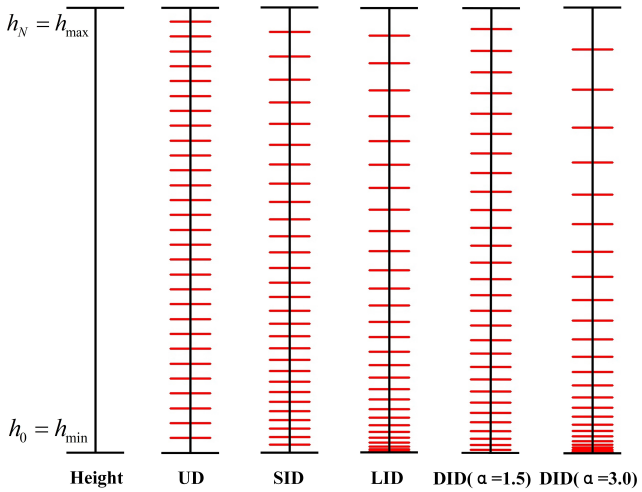

Figure 8. **Height Discretization Methods.** Height h_i is discretized over a height range $[h_{min}, h_{max}]$ into N discrete bins. From left to right, these are uniform discretization (UD), spacing-increasing discretization (SID), linear-increasing discretization (LID) and adjustable dynamic-increasing discretization(DID). For the dynamic-increasing discretization (DID) strategy, height bins with large α are more densely distributed when approaching the h_{min} than the small hyper-parameter α conditions.

Table 7. **Ablating our dynamic discretization on DAIR-V2X-I dataset.** Compared to the uniform discretization(UD), our method achieves on average 1% improvement in average precision.

Spacing		Veh.$_{(IoU=0.5)}$			Ped.$_{(IoU=0.25)}$			Cyc.$_{(IoU=0.25)}$		
DID (α)	UD	Easy	Mid	Hard	Easy	Mid	Hard	Easy	Mid	Hard
	✓	75.63	63.75	63.85	25.82	25.47	25.35	47.52	47.47	47.19
✓(1.5)		76.24	64.54	64.13	26.47	25.79	25.72	48.55	48.21	47.96
✓(2.0)		**76.61**	**64.71**	**64.76**	**27.34**	**26.09**	**25.33**	**49.68**	**48.84**	**48.58**

Table 8. **Results with point cloud supervision on DAIR-V2X-I dataset.** We can observe that for both BEVDepth and BEVHeight, LiDAR point cloud supervision did not help in terms of evaluation results. This is another evidence that road-side perception is different from the ego-vehicle one.

Method	Veh.$_{(IoU=0.5)}$			Ped.$_{(IoU=0.25)}$			Cyc.$_{(IoU=0.25)}$		
	Easy	Mid	Hard	Easy	Mid	Hard	Easy	Mid	Hard
BEVDepth	71.56	60.75	60.85	21.55	20.51	20.75	40.83	40.66	40.26
BEVDepth†	71.09	60.37	60.46	21.23	20.84	20.85	40.54	40.34	40.32
BEVHeight	75.58	63.49	63.59	26.93	25.47	**25.78**	47.97	47.45	48.12
BEVHeight†	**75.64**	**63.61**	**63.72**	**27.01**	**25.55**	25.34	**48.03**	**47.62**	**48.19**

† denotes training with PointCloud supervision.

than that without supervision. As for our BEVHeight, although there is a slight improvement under the IoU=0.5 condition, the overall gain is not apparent. This can be explained by the fact that the background in roadside scenarios is stable. These background point clouds fail to provide adequate supervised information and increase the difficulty of model fitting.

A.6. Results on V2X-Sim Dataset

To certify the effectiveness of our method in multi-view scenarios, we conduct experiments on V2X-Sim [4] simulation dataset that contains four surround roadside cameras. As shown in Tab. 9, our BEVHeight surpass the BEVDepth by more than 10.88%, 21.15% on vehicle and cyclist respectively, which verifies the effectiveness of our method.

Table 9. **Comparison on the V2X-Sim Detection Benchmark.**

Method	Vehicle$_{(IoU=0.5)}$			Cyclist$_{(IoU=0.25)}$		
	Easy	Mod.	Hard	Easy	Mod.	Hard
BEVDepth [3]	81.99	81.39	81.31	45.95	45.93	45.90
BEVHeight	92.80	92.27	92.15	67.24	67.08	67.00

A.7. Effectiveness on Multi Depth-based Detectors

We extend our modules on BEVDepth [3] and BEVDet [2] on DAIR-V2X-I [6] and present the results here. Replacing the depth-based projection in BEVDepth [3], our method achieves a performance increase of 2.19%, 5.87%, 4.61% on vehicle, pedestrian and cyclist. Similarly, our approach surpasses the origin BEVDet by 8.56%, 5.35%, 8.60% respectively.

Table 10. **Ablation studies on different depth-based methods.** Here, we conduct the evaluation on DAIR-V2X-I val set, and report the results of three types of objects, vehicle (veh.), pedestrian (ped.) and cyclist (cyc.).

Method	VT	Veh.$_{(IoU=0.5)}$			Ped.$_{(IoU=0.25)}$			Cyc.$_{(IoU=0.25)}$		
		Easy	Mod.	Hard	Easy	Mod.	Hard	Easy	Mod.	Hard
BEVDepth [3]	D	75.50	63.58	63.67	34.95	33.42	33.27	55.67	55.47	55.34
	H	77.78	65.77	65.85	41.22	39.29	39.46	60.23	60.08	60.54
BEVDet [2]	D	59.59	51.92	51.81	12.61	12.43	12.37	34.91	34.32	34.21
	H	69.42	60.48	59.68	18.11	17.81	17.74	44.69	42.92	42.34

VT denotes view transformation, D,H represents depth-based and height-based ones.

A.8. More Results on DAIR-V2X-I Dataset

Tab. 11 shows the experimental results of deploying our proposed approach on the DAIR-V2X-I [6] val set. Under the same configurations (e.g., backbone and BEV resolution), our model outperforms the BEVDepth [3] baselines by a large marge, which demonstrates the admirable performance of our approach.

A.9. More Visualizations

In Fig. 9 and Fig. 10, we show more visualization results on the DAIR-V2X-I [6] dataset. As can be seen from the samples in I/II-(a) clean, our BEVHeight manage to detect objects in middle and long-distances. As for the extrinsic disturbance cases in I/II-(b) and I/II-(c), our method can still guarantee the detection accuracy in terms of cars, pedestrian

Table 11. **Comparison on the DAIR-V2X-I Detection Benchmark.** Here, we report the results of three types of objects: Vehicle, Pedestrian and Cyclist. Each object is categorized into three settings according to the difficulty defined in [6]. Our BEVHeight manages to surpass the BEVDepth baseeline over a margin of 2% - 6% under the same configurations.

Method	Scale of Detector		AP3D								
			Vehicle$_{(IoU=0.5)}$			Pedestrian$_{(IoU=0.25)}$			Cyclist$_{(IoU=0.25)}$		
	Backbone	BEV	Easy	Middle	Hard	Easy	Middle	Hard	Easy	Middle	Hard
BEVDepth [3]	R50	128x128	73.05	61.32	61.19	22.10	21.57	21.11	42.85	42.26	42.09
BEVDepth [3]	R101	128x128	74.81	62.44	62.31	24.49	23.33	23.17	44.93	44.02	43.84
BEVDepth [3]	R101	256x256	75.50	63.58	63.67	34.95	33.42	33.27	55.67	55.47	55.34
BEVHeight	R50	128x128	76.61	64.71	64.76	27.34	26.09	26.33	49.68	48.84	48.58
BEVHeight	R101	128x128	76.93	64.97	65.03	28.53	27.15	27.48	51.39	50.83	50.44
BEVHeight	R101	256x256	**77.78**	**65.77**	**65.85**	**41.22**	**39.29**	**39.46**	**60.23**	**60.08**	**60.54**

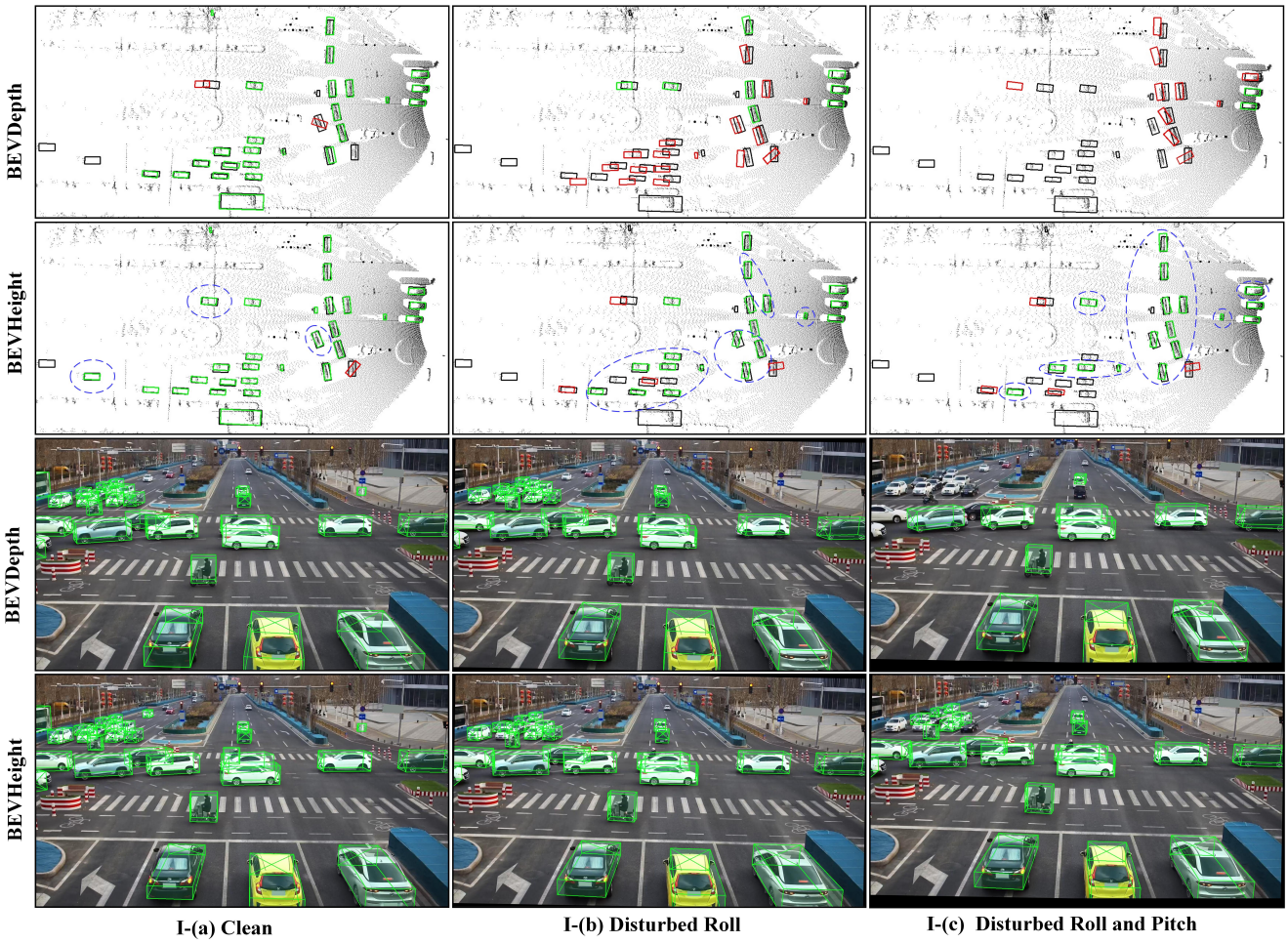

Figure 9. **Visualization Results of BEVDepth and our proposed BEVHeight under the extrinsic disturbance.** We use boxes in **red** to represent false positives, **green** boxes for truth positives, and **black** for the ground truth. The truth positives are defined as the predictions with IoU>0.5 for vehicle and IoU>0.25 for pedestrian and cyclist. I/II-(a) Clean means the original image without any processing; I/II-(b) Disturbed Roll denotes camera rotate 1 degree along roll direction; I/II-(c) Disturbed Roll and Pitch represents camera rotate 1 degree along roll and pitch directions simultaneously. We use **blue** dashed ovals to highlight the pronounced improvements in predictions.

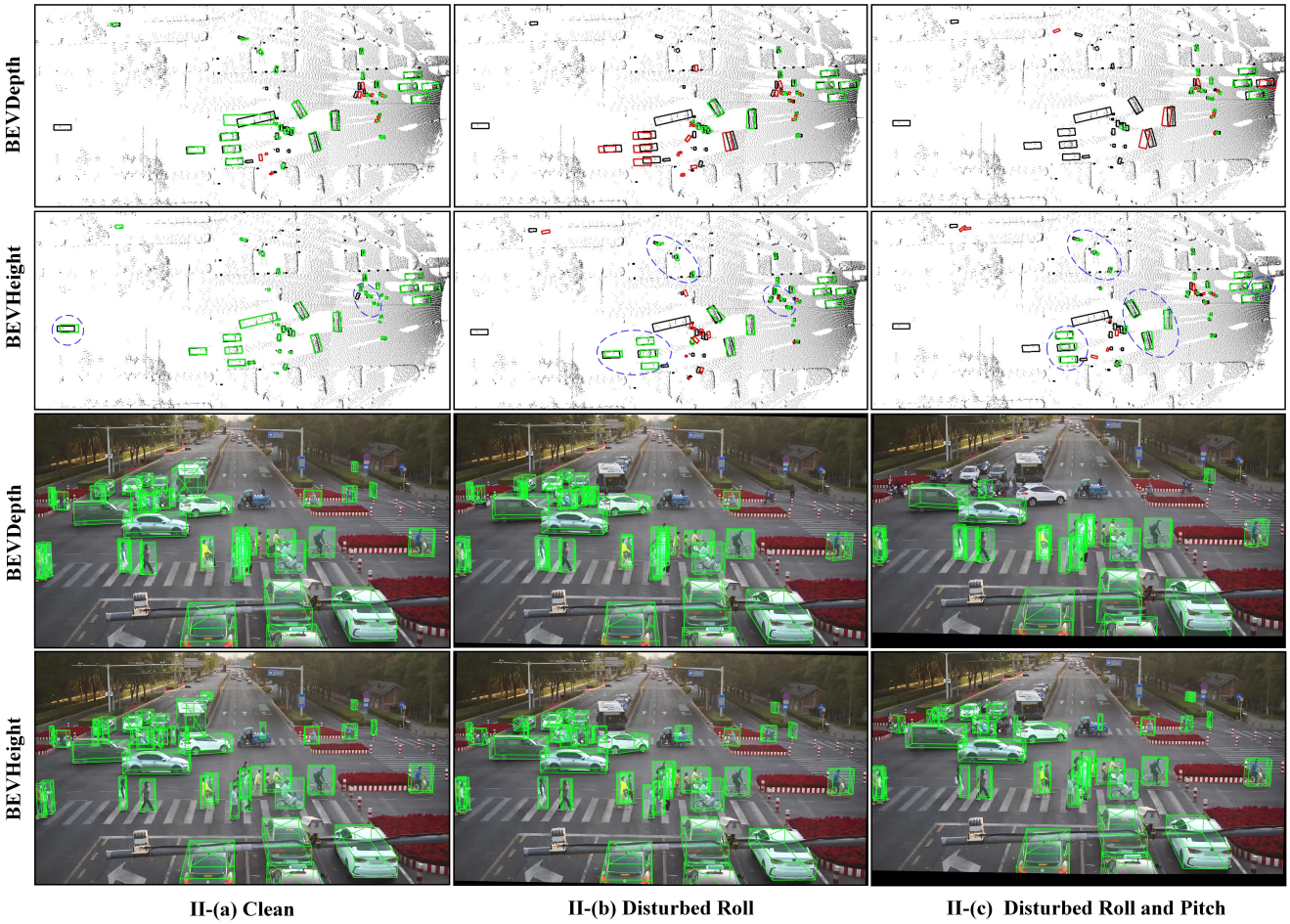

Figure 10. **Visualization Results of BEVDepth and our proposed BEVHeight under the extrinsic disturbance in another scene.**

and cyclist. It can be concluded that our method can significantly improve the accuracy in middle and long-distances and the robustness to extrinsic disturbance.

References

[1] Huan Fu, Mingming Gong, Chaohui Wang, Kayhan Batmanghelich, and Dacheng Tao. Deep ordinal regression network for monocular depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018. 1

[2] Junjie Huang, Guan Huang, Zheng Zhu, and Dalong Du. Bevdet: High-performance multi-camera 3d object detection in bird-eye-view. *arXiv preprint arXiv:2112.11790*, 2021. 2

[3] Yinhao Li, Zheng Ge, Guanyi Yu, Jinrong Yang, Zengran Wang, Yukang Shi, Jianjian Sun, and Zeming Li. Bevdepth: Acquisition of reliable depth for multi-view 3d object detection. *arXiv preprint arXiv:2206.10092*, 2022. 1, 2, 3

[4] Yiming Li, Dekun Ma, Ziyan An, Zixun Wang, Yiqi Zhong, Siheng Chen, and Chen Feng. V2x-sim: Multi-agent collaborative perception dataset and benchmark for autonomous driving. *IEEE Robotics and Automation Letters*, 7(4):10914–10921, 2022. 2

[5] Yunlei Tang, Sebastian Dorn, and Chiragkumar Savani. Center3d: Center-based monocular 3d object detection with joint depth understanding. *arXiv: Computer Vision and Pattern Recognition*, 2020. 1

[6] Haibao Yu, Yizhen Luo, Mao Shu, Yiyi Huo, Zebang Yang, Yifeng Shi, Zhenglong Guo, Hanyu Li, Xing Hu, Jirui Yuan, et al. Dair-v2x: A large-scale dataset for vehicle-infrastructure cooperative 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21361–21370, 2022. 2, 3